Small vs. Big Data in Language Research: Challenges and Opportunities

A. Seza Doruz
Independent Researcher
a.s.dogruoz@gmail.com

Abstract

Mobile communication tools and platforms provide various opportunities for users to interact over social media. With the recent developments in computational research and machine learning, it has become possible to analyze large chunks of language related data automatically and fast. However, these tools are not readily available to handle data in all languages and there are also challenges handling social media data. Even when these issues are resolved, asking the right research question to the right set and amount of data becomes crucially important.

Both qualitative and quantitative methods have attracted respectable researchers in language related areas of research. When tackling similar research problems, there is need for both top-down and bottom-up data-based approaches to reach a solution. Sometimes, this solution is hidden under an in-depth analysis of a small data set and sometimes it is revealed only through analyzing and experimenting with large amounts of data. However, in most cases, there is need for linking the findings of small data sets to understand the bigger picture revealed through patterns in large sets.

Having worked with both small and large language related data in various forms, I will compare pros and cons of working with both types of data across media and contexts and share my own experiences with highlights and lowlights.

Keywords: social media data, machine learning, small vs. large data sets, multilingualism

References

Nguyen, D. and Doğruöz, A. (2014). Word level language identification in online multilingual communication. In *EMNLP*.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational linguistics*.

Papalexakis, E. and Doğruöz, A. S. (2015). Understanding multilingual social networks in online immigrant communities. In *Proceedings of the 24th International Conference on World Wide Web*, pages 865–870. ACM.